Sussex Research Online

Demographic information prediction: a portrait of smartphone application Users

Article (Accepted Version)

Qin, Zhen, Wang, Yilei, Cheng, Hongrong, Zhou, Yingjie, Sheng, Zhengguo and Leung, Victor C M (2016) Demographic information prediction: a portrait of smartphone application Users. IEEE Transactions on Emerging Topics in Computing (99). ISSN 2168-6750

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/60886/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk

Demographic Information Prediction: A Portrait of Smartphone Application Users

Zhen Qin*, Yilei Wang*, Hongrong Cheng*, Yingjie Zhou†*‡, Zhengguo Sheng§ and Victor C.M. Leung¶
*University of Electronic Science and Technology of China, Chengdu, China
†Sichuan University, Chengdu, China
§University of Sussex, United Kingdom
¶University of British Columbia, Canada

Abstract—Demographic information is usually treated as private data (*e.g.,* gender and age), but has been shown great values in personalized services, advertisement, behavior study and other aspects. In this paper, we propose a novel approach to make efficient demographic prediction based on smartphone application usage. Specifically, we firstly consider to characterize the data set by building a matrix to correlate users with types of categories from the log file of smartphone applications. Then, by considering the category-unbalance problem, we make use of the correlation between users' demographic information and their requested Internet resources to make the prediction, and propose an optimal method to further smooth the obtained results with category neighbors and user neighbors. The evaluation is supplemented by the dataset from real world workload. The results show advantages of the proposed prediction approach compared with baseline prediction. In particular, the proposed approach can achieve 81.21% of Accuracy in gender prediction. While in dealing with a more challenging multi-class problem, the proposed approach can still achieve good performance (*e.g.,* 73.84% of Accuracy in the prediction of age group and 66.42% of Accuracy in the prediction of phone level).

Index Terms—demographic information, personalized, prediction, smartphone application users.

1 INTRODUCTION

Recent technology trends in industry and academia are seeking practice solutions to provide personalized services and improve user experience. For example, Google provides personalized search results, where different users searching for the same terms may receive different responses [1]. Targeted advertising is another example as it recommends proper users to advertisers upon their behaviors on the Internet [2]. As reported by the recent study [3], the targeted advertising is more effective than the ordinary advertisement.

A recent study has investigated users' resource consumption in mobile Internet to characterize user behavior [4]. To provide personalized services or targeted advertising, it is also necessary to study the interests of potential customers, where their personal data (*e.g.,* browsing histories, search interests, geographic information and even demographic information) play an important role in user behavior study. Demographic information has been shown great value to provide accurate personalized services. However, it is a kind of personal data that users prefer keeping in private, and usually not easy to obtain from public.

With the development of network-enabled smartphones and networking technology, smartphones have changed our lifestyles and linked the virtual network world and the actual social world. Nowadays, people use a variety of smartphone applications in everyday life. While users might have different interests, the applications installed on each smartphone could be distinct (*e.g.,* females prefer "fashion related" applications, while males prefer "sports related" applications). Even for the same application, its usage (*e.g.,* the "queries" when using the search engine application published by Google) by each user could be varied. Since the smartphone and its user are tightly correlated, the smartphone application usage could be used to predict personal demographic information.

In this paper, we attempt to infer their demographic information (*e.g.,* gender, age group and phone level) by leveraging the difference of individual behaviors on using their smartphone applications - to make a portrait of smartphone application users. Specifically, our proposed prediction method can be processed as follows: (i) We match the entries in smartphone applications log with types of categories according to their required Internet resources, and build a matrix to correlate users and categories. (ii) We provide the evaluation metrics and baseline performance of our study and verify that the duration of our dataset is sufficient for making stable prediction. (iii) We then consider the category-unbalance problem and leverage Bayes-based classifier to predict users' demographic information from the matrix. (iv) By combining the information from category neighbors and user neighbors, we smooth and optimize the prediction results. (v) Finally, the performance of the proposed approach can be evaluated based on a dataset from real collected workload.

The paper is organized as follows. In Section 2, we provide an overview of the related work in demographic information prediction. The objective and dataset processing is introduced in Section 3. In Section 4, based on our dataset, we investigate the required duration and user amount of training data for stable prediction, and provide the baseline

‡*This is the corresponding author. Email: yjzhou09@gmail.com; yjzhou@scu.edu.cn.*

prediction for our study. The proposed prediction approach is presented in Section 5. Experimental results to evaluate the approach are shown in Section 6. In Section 8, we conclude the paper and state our future work.

2 Related Work

Demographic information has attracted increasing interests recently and a number of distinct prediction methods have been proposed to enrich personalized services and application.

Prediction based on user-generated content. User-generated content (*e.g.,* profiles, blogs, comments, photos, videos) is everywhere on today's Internet. Demographic information is usually included in profiles. Nevertheless, only a limited number of users allow public access to it. Beyond profiles, other types of content can also be used to infer users' demographic information. Garera et al. [5] proposed that demographic information can be predicted by analyzing users' writing and speaking styles. Yan et al. [6] investigated techniques using blog content to infer user gender, while Burger et al. [7] identified the gender of Twitter users by leveraging their tweets. Kelly et al. [8] leveraged the smartphone data to investigate the relation between location behavior patterns and particular demographic and social characteristics about an individual. Other researchers leveraged or combined different types of user-generated content to make the demographic prediction [9] [10].

Prediction based on users' Internet activity histories. Internet activity histories include searching histories, browsing histories, etc. There have been state-of-arts works on searching and browsing histories. Ingmar et al. [11] investigated the search behaviors of different demographic groups, and verified that the demographic description of search engine users agrees well with the distribution of US population. Another work of Ingmar [12] has shown that certain types of search queries can be related to distinct demographic groups. Based on the browsing histories of certain users whose demographic information is known, Hu et al. [13] inferred the gender and age of other users who browse the same webpages. Kabbur et al. [14] and Goel et al. [15] used machine learning techniques to predict demographic information by analyzing the HTMLs, key words and hyperlinks in user browsing histories. Suranga et al. [16] leveraged the difference in the snapshots of smartphone applications installed by users with multicultural background, and predicted user traits about origin, native language, marital status, religion and whether they have any child.

In comparison, by leveraging the smartphone application usage, we propose a novel approach to predict demographic information of users whom have the similar culture background. The logs of smartphone applications record the Internet activity histories with particular characteristics. For a specific Internet query or webpage, the topic is usually unambiguous that can be associated with a certain demographic group, but the story is totally different when it comes to the topic for a specific smartphone application usage (*e.g.,* for a video application, adults might watch daily news, while children could choose a cartoon). Thus, our prediction approach, introduced in the next section, is going to provide a solution to the particular characteristics. This paper is an extension of the work from [17] that studied users' demographic information based on their smartphone applications logs.

3 Objective and Dataset

3.1 Objective Definition

Before introducing our approach for demographic prediction, we firstly define our objective. Given some users are with their demographic information, the objective of this paper is to provide an approach to infer the demographic information of other users based on their usage of smartphone applications.

The demographic information we mainly consider in this paper are gender, phone level and age group. The gender prediction is defined as classifying users as male or female. As shown in Table 1, the phone level prediction is defined as classifying users' phone price into one of the price range [18], while the age group prediction is defined as classifying users into one of the age range [19].

TABLE 1
Phone Levels & Age Groups

Phone Level	Level1	Level2	Level2	\	
Price Range	<$150	$150-300	>$300		
Age Group	Teenage	Youngster	Young	Mid-age	Elder
Age Range	<18	18-24	25-44	45-59	>59

A direct way for prediction is to train a classifier from the aspect of applications, and correlate different demographic groups with certain smartphone applications. Suranga et al. [16] use Support Vector Machine (SVM) classifier [20] to predicted user traits from a snapshot of applications installed on a smartphone. Their prediction about user traits mainly focused on the origin, native language, marital status, religion and whether they have any child. In this paper, the user demographic information we predicted is a more challenging problem. As the users in our dataset are from a province of mainland China (rather than a multicultural community as the users in [16] are), they have the similar culture and share the similar traits (*e.g.,* origin, native language and religion) which is quite common for non-immigrant country. Under such scenario, it requires a precise data feature extraction before training a classifier. To verify this, we have trained the SVM classifier used in [16] and three other classifiers (*e.g.,* Logistic [21], Naive Bayes (NB) [22], and Decision Tree(DT) [23]). The results are very poor. Even for the binary classification - gender prediction, both accuracy and F1 value are no more than 60%.

To solve the problem, we consider the fact that users with different demographic information might use a same application, and think that training a classifier from the aspect of applications is coarse-grained. It is necessary to extract some fine-grained data features before training a classifier. Fortunately, for an application, its usage can be related to different kinds of topics (*e.g.,* a video play application can be related to sports, tourism, cartoon, etc.) that indicate the users' interests. In this way, by adding up the interest weights of different applications, we correlate different demographic groups with fine-grained users' interests as below.

TABLE 2
Demographic Distribution

	Age Groups					Total
	Teen age	Youngster	Young	Mid-age	Elder	
Male-Level1	0.52%	2.90%	7.56%	14.34%	6.56%	31.88%
Male-Level2	0.75%	2.90%	5.89%	10.54%	3.04%	23.12%
Male-Level3	0.10%	0.70%	2.46%	4.41%	1.27%	8.94%
Female-Level1	0.36%	1.81%	4.66%	8.13%	2.85%	17.81%
Female-Level2	0.60%	2.02%	3.80%	5.79%	1.42%	13.63%
Female-Level3	0.08%	0.51%	1.37%	2.14%	0.52%	4.62%
Total	2.41%	10.84%	25.74%	45.35%	15.66%	100%

3.2 Dataset Processing

The dataset we used is real world smartphone applications workload obtained from a network service provider for a period of four months from October 2014 to January 2015 including workdays and weekends. The dataset includes 32,660 users, whose demographic attributes distribution is shown in Table 2. The users are volunteers of the study. To protect the privacy of users, their ids are anonymous.

The dataset covers 438 unique smartphone applications and records the logs of such applications, which are the amount of entries when smartphone applications fetch resources from the Internet. 179,954,181 entries are included in the dataset. Each entry includes a numerical code which is generated from user ID and corresponding Internet resource. By using Regular Expression Matching with different key words (*e.g.,* "football" refers to a category of "sports"), the corresponding Internet resource is matched to a certain category. Thus, each entry can be used to match a numerical code (generated from user ID) and a category.

Then, we define a topic as a category of users' interests. The dataset can be modeled as a weighted graph $G = (V, E)$, where a node in V represents a user or a category, and an edge in E represents the frequency of a user's request to a category. The nodes in V can be divided into two subsets, $U = u_1, u_2, ..., u_m$ and $C = c_1, c_2, ..., c_n$, where U represents the users and C represents the categories.

Next, the graph G can be represented by an adjacent matrix R, whose element r_{ij} is the frequency of user u_i's request to category c_j. In our approach, the frequency is recorded as the *request times* in a given period. In this way, the adjacent matrix R correlates users and categories. Now, we can train the classifies from the aspect of categories.

4 Baseline Prediction

In this section, we provide the evaluation metrics and the prediction results based on the training of four classifiers (*e.g.,* Logistic, Naive Bayes(NB),Decision Tree(DT) and Support Vector Machine(SVM)) [20], [21], [22], [23]. Moreover, by increasing the duration of training data, we investigate how long the duration of training data is required for stable prediction. Similarly, by increasing the user amount of training data, we investigate how much the user amount of training data is required for stable prediction.

4.1 Evaluation Metric

The performance of prediction is evaluated by Accuracy (Acc), Precision $(Prec)$, Recall (Rec) and F1 value $(F1)$ [24]. In an information retrieval system, the predicted examples are classified as true positives (tp), false positives (fp), false negatives (fn) and true negatives (tn).

Acc is the proportion of correctly predicted examples in the set of all examples, and it is defined as Formula 1:

$$Acc = \frac{tp + tn}{tp + fp + fn + tn} \tag{1}$$

$Prec$ is the proportion of correctly predicted examples in the set of all examples assigned to the target class, and it is defined as Formula 2:

$$Prec = \frac{tp}{tp + fp} \tag{2}$$

Rec is the proportion of correctly predicted examples out the set of all examples having the target class, and it is defined as Formula 3:

$$Rec = \frac{tp}{tp + fn} \tag{3}$$

It is certainly that the higher both $Prec$ and Rec are, the better the performance is. However, high $Prec$ usually demands strict conditions which would lead low Rec. As $Prec$ and Rec are sometimes contradictory, $F1$ value is a trade off between $Prec$ and Rec [25], and is defined as a combination value of $Prec$ and Rec in Formula 4:

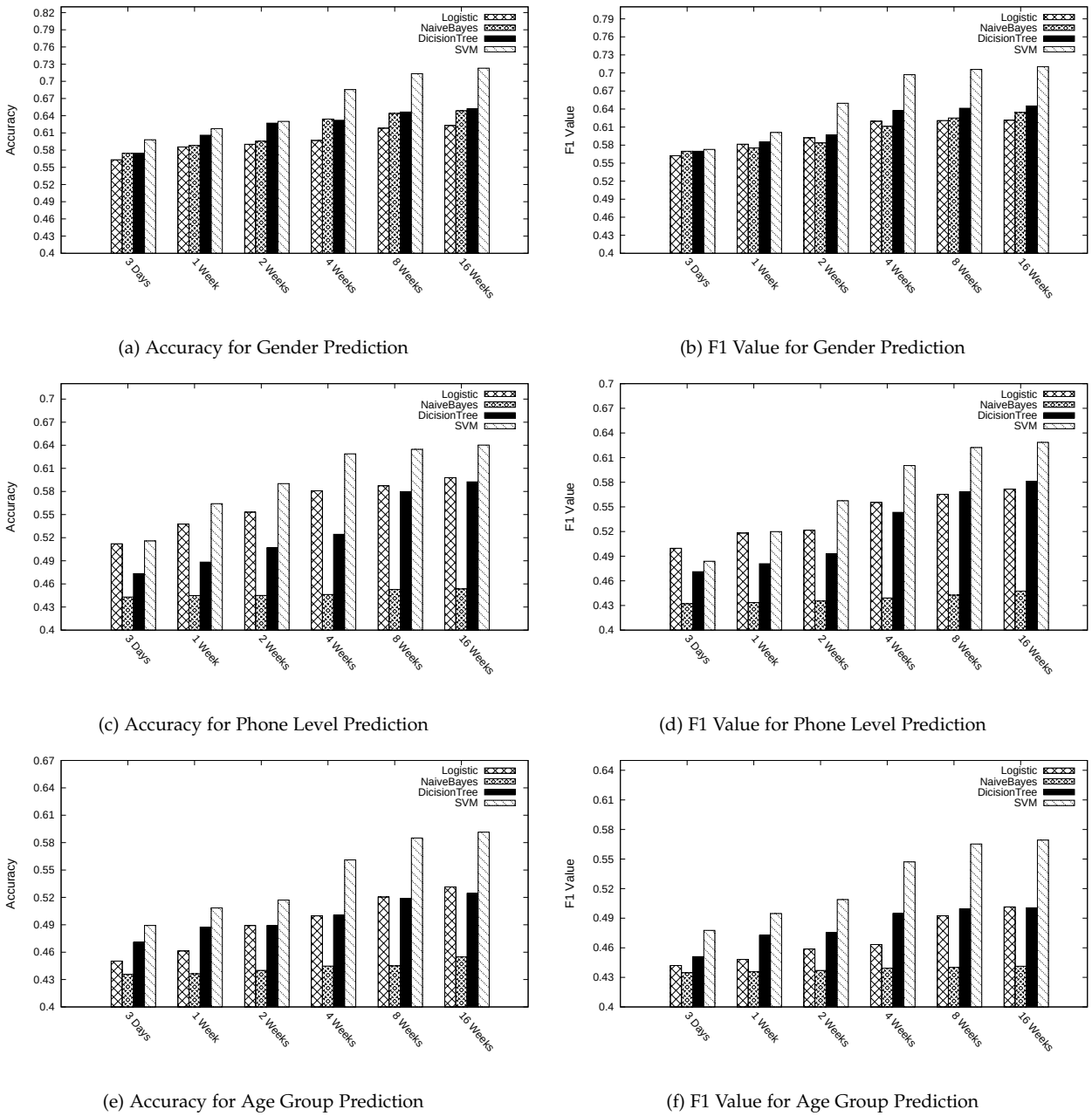
$$F1 = \frac{2Prec \times Rec}{Prec + Rec} \tag{4}$$

In this paper, we apply Acc and $F1$ to get performance values in the evaluation.

4.2 Required Data Duration for Stable Prediction

Before making the demographic prediction, it is necessary to investigate the required data duration for achieving stable prediction results. In particular, we trained classifiers with fixed user amount (*e.g.,* 32,000 users) but different duration of dataset (*e.g.,* 3 days, 1 week, 2 weeks, etc) to find a stable status for predicting users' demographic information. The experimental results are shown in Fig. 1. The results are obtained by training the four classical classifiers. Fig. 1(a), Fig. 1(c) and Fig. 1(e) show the Accuracy of prediction in different duration of dataset, while Fig. 1(b), Fig. 1(d) and Fig. 1(f) show the $F1$ value of prediction in different duration of dataset. It can be found that both Acc and $F1$ received better performance with the increasing of the duration of training data. Moreover, between the duration of 3 days and 8 weeks, there are obvious performance improvement in both Acc and $F1$ for all four classifiers.

Most importantly, it should be noticed that the prediction results can achieve stable performance (or convergent results) with the duration of more than 8 weeks. Based on our dataset, this means that if users' smartphone applications usage were traced more than 8 weeks, their demographic information would be well predicted. Thus, the duration of our dataset stated in above section is sufficient for making stable prediction, as it lasts four months.



(a) Accuracy for Gender Prediction

(b) F1 Value for Gender Prediction

(c) Accuracy for Phone Level Prediction

(d) F1 Value for Phone Level Prediction

(e) Accuracy for Age Group Prediction

(f) F1 Value for Age Group Prediction

Fig. 1. Prediction in Different Data Duration

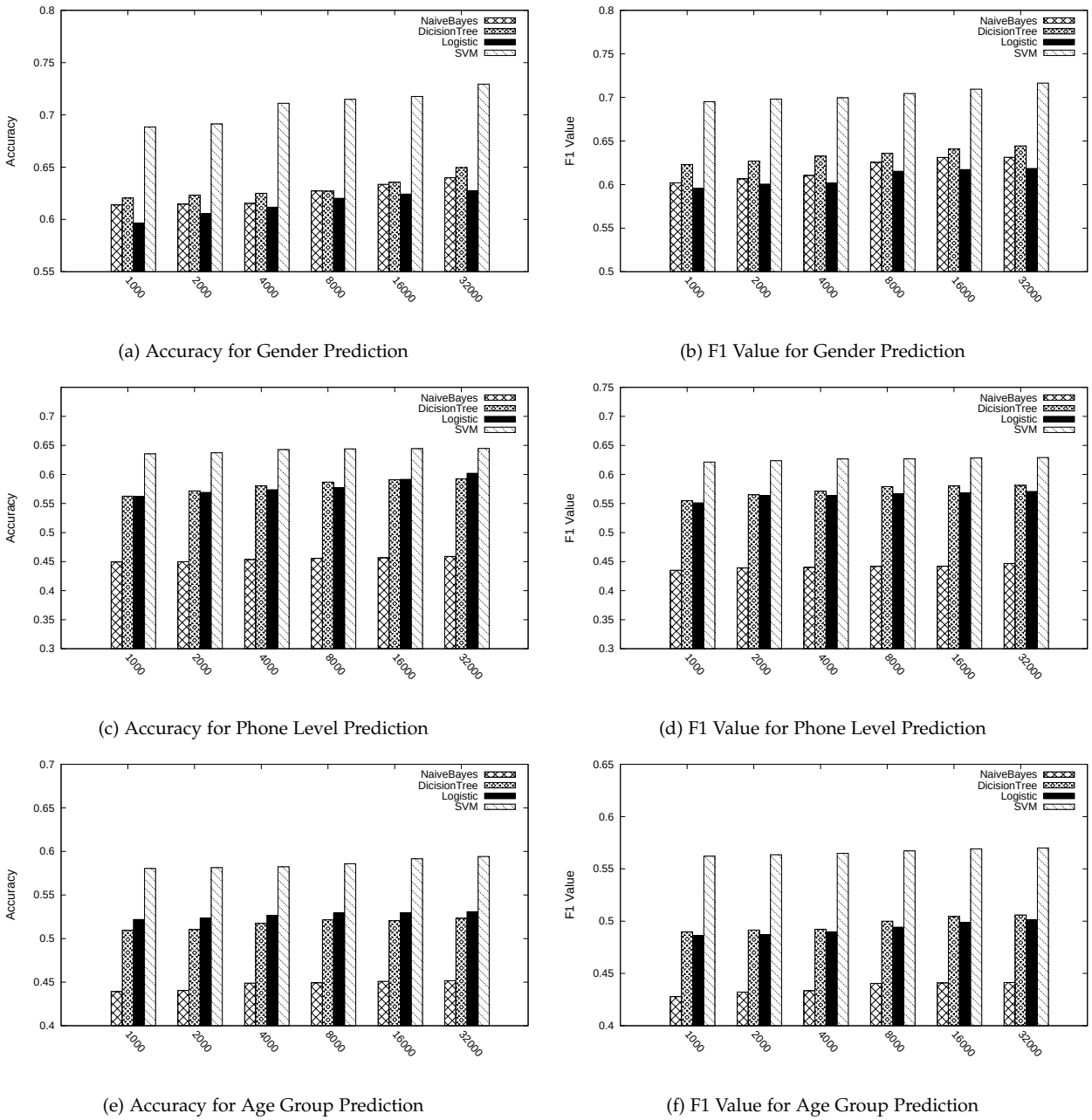
4.3 Required User Amount for Stable Prediction

Similarly, it is also necessary to investigate the required user amount for achieving stable prediction results. In particular, by setting the data duration as 16 weeks, we train classifiers with different user amount of dataset (*e.g.,* 1,000, 2,000, 4,000, etc) to find a stable status for predicting users' demographic information. The experimental results are shown in Fig. 2. The results are obtained by training the four classical classifiers. Fig. 2(a), Fig. 2(c) and Fig. 2(e) show the Accuracy of prediction in different user amount of dataset, while Fig. 2(b), Fig. 2(d) and Fig. 2(d) show the $F1$ value of prediction in different user amount of dataset. It can be found that both Acc and $F1$ received better performance with the increasing of user amount of training data. Moreover, between the user amount of 1000 and 4000, It can be noticed that the performance are also improved in both Acc and $F1$ for all four classifiers.

When the user amount varies from 1,000 to 32,000, it should be noticed that the Accuracy of prediction can achieve stable performance with the user amount of more than 4,000, while the $F1$ value of prediction keeps stable when the user amount varies from 1,000 to 32,000. Thus, in our dataset, the amount of 32,660 users is sufficient for making stable prediction.

4.4 Prediction Results of Classical Classifiers

Here, we provide the prediction results of four classical classifiers as baselines. In the experiments, 10-fold cross-validation was deployed for our dataset, the duration of training data is 16 weeks, and the user amount of training data is 32,000. Thus, the baselines can achieve stable performance.



(a) Accuracy for Gender Prediction

(b) F1 Value for Gender Prediction

(c) Accuracy for Phone Level Prediction

(d) F1 Value for Phone Level Prediction

(e) Accuracy for Age Group Prediction

(f) F1 Value for Age Group Prediction

Fig. 2. Prediction in Different User Amount

The prediction results are also shown in above six figures. The gender prediction results are 62.26% and 62.13% for Logistic, 64.85% and 63.41% for NB, 65.23% and 64.51% for DT, 72.31% and 71.05% for SVM, in term of Acc and $F1$, respectively. For the prediction of users' phone level, the values of Acc and $F1$ are 59.81% and 57.16% for Logistic, 45.34% and 44.75% for NB, 59.25% and 58.12% for DT, 64.03% and 62.85% for SVM, respectively. While predicting users' age group, the results are 53.14% and 50.15% for Logistic, 45.47% and 44.15% for NB, 52.43% and 50.05% for DT, 59.14% and 56.93% for SVM, in term of Acc and $F1$, respectively. It can be found that Logistic, NB and DT have the similar performance for gender prediction, while Logistic and DT outperform NB in the prediction of phone level and age group. Moreover, SVM receives the best prediction performance in the four classical classifiers. These results are the baselines of our study, a novel prediction approach will be introduced in the following section.

5 THE PROPOSED PREDICTION APPROACH

The proposed prediction approach is stated in this section. In particular, by analyzing the training data, we notice that some categories are over weighted thus would bias prediction results. Moreover, we argue that people with similar demographic attributes may have similar interests. It means that the prediction results can be optimized with user neighbors and categories neighbors. Thus, by dealing with category-unbalance and optimizing the prediction results with neighbors' information, we propose a novel prediction approach by leveraging a Bayes-based classifier.

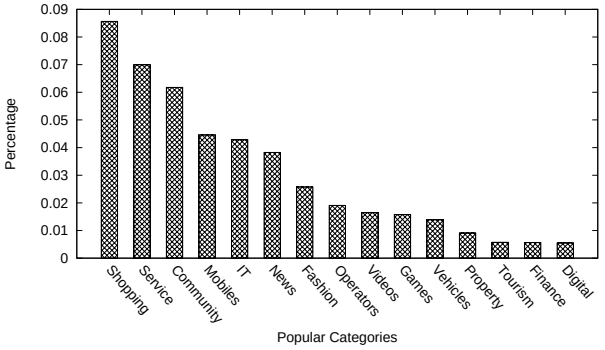


Fig. 3. Users interest distribution

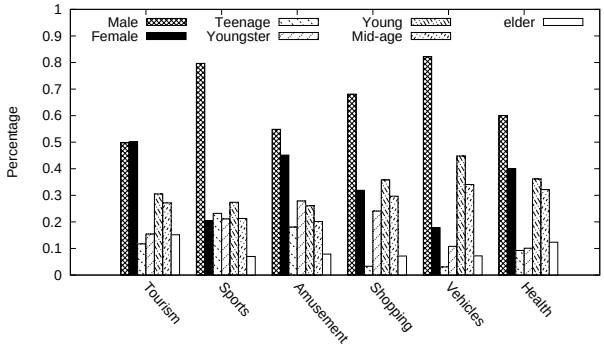


Fig. 4. The Perspectives of Smartphone Application Usage

5.1 Category-Unbalance Problem

For each category in the adjacent matrix R (stated in Section 3), it is feasible to compute its demographic distribution (*e.g.,* the distribution of male and female), which is going to be used as *prior probability* in Bayes-based classifier. Ideally, all the categories are considered to have the same weight. However, the *request times* for a category can have orders of magnitude larger than the *request times* for another category (*e.g.,* shopping *vs.* tourism as shown in Fig. 3). For the categories with less *request times* (*e.g.,* vehicles and health), they can also reflect the difference in demographic distribution as shown in Fig. 4. Thus, such categories should be assigned a higher weight. If we directly import the adjacent matrix R into a Bayes-based classifier, some categories with more *request times* would be over weighted, and some categories with less *request times* would be under weighted, thus bias the prediction results.

For a category, if its *request times* is more for a user group but less for others, it is a good feature for classification. This logical way of thinking follows the idea of Term Frequency - Inverse Document Frequency (TF-IDF) [26] [27], which is often used as a weighting factor in information retrieval and text mining. To reduce the adverse influence of over/under weighted categories, the idea of TF-IDF is leveraged to process the adjacent matrix R and assign appropriate weights for categories. The adjacent matrix R is normalized as follows:

$$TF(r_{ij}) = \frac{r_{ij}}{\sum_{j=1}^{n} r_{ij}} \quad (5)$$

$$IDF(r_{ij}) = log\frac{m}{1+\sum_{i=1}^{m} I(r_{ij})} \quad (6)$$

$$TFIDF(r_{ij}) = TF(r_{ij}) * IDF(r_{ij}) \quad (7)$$

where, edge r_{ij} is the *request time* between the i^{th} user and the j^{th} category, m is the number of users. $I(r_{ij})$ is an indicator function. If r_{ij} has a non-zero value, $I(r_{ij})$ will be 1, otherwise $I(r_{ij})$ will be 0. Thus, we obtain a new adjacent matrix $R^{'}$, whose element r'_{ij} is $TFIDF(r_{ij})$, to correlate users and categories.

5.2 Compute Prior Probability

Our prediction approach is based on the perspectives of smartphone application usage. Fig. 4 shows the statistics of some categories with different demographic groups. It can be noticed that males were more interested in the categories of sports and vehicles, which can be leveraged to predict users' gender. Moreover, in Fig. 4, elder people paid more attention for tourism than other categories. Even the *request time* shown in Fig. 3 is low, the category of tourism can still be used to predict users' age group. Thus, it proves the necessity to calculate the nomalized adjacent matrix $R^{'}$.

Now, we compute the prior probability of every demographic attribute for each category. Given the smartphone application usage of some users with their demographic information, we can obtain the new adjacent matrix $R^{'}$ by analyzing the smartphone application usage. Combining the matrix $R^{'}$ with these training users' demographic information, the prior probability of categories are computed as follows:

$$\Pr(c_j|A) = \frac{\sum_{i=1}^{m} r'_{ij} u_i(A) + 1}{\sum_{i=1}^{m}\sum_{j=1}^{n} r'_{ij} u_i(A) + n} \quad (8)$$

where, $Pr(c_j|A)$ is the prior probability of a demographic attribute A (*e.g.,* male or female) for category c_j, r'_{ij} is the element in matrix $R^{'}$, m is the number of users, n is the number of categories, and $u_i(A)$ is an indicator function. If attribute A is true for the i^{th} user, $u_i(A)$ will be 1, otherwise $u_i(A)$ will be 0.

5.3 Predict Users Demographic Information

Based on the prior probabilities computed in above subsection, here we use a Bayes-based classifier [28] to predict users' demographic attributes. Assuming the categories are independent, the probability of user u_i's demographic attribute A is computed as follows:

$$\begin{aligned} \Pr(A|u_i) &\propto Pr(A|\{categories\}) \\ &= \frac{Pr(\{categories\}|A)Pr(A)}{Pr(\{categories\})} \\ &= \frac{\prod_j^k Pr(c_j|A)^{r'_{ij}} Pr(A)}{Pr(\{categories\})} \end{aligned} \quad (9)$$

where, A is a demographic attribute, $\{categories\}$ is the set of categories that user u_i correlated with, $Pr(\{categories\}|A)$ is the set of prior probabilities, and $Pr(\{categories\})$ can be offset by normalization.

5.4 Optimize Prediction Results

As people with similar demographic attributes may have similar interests, it is possible to optimize prediction results by leveraging the demographic attributes of other users

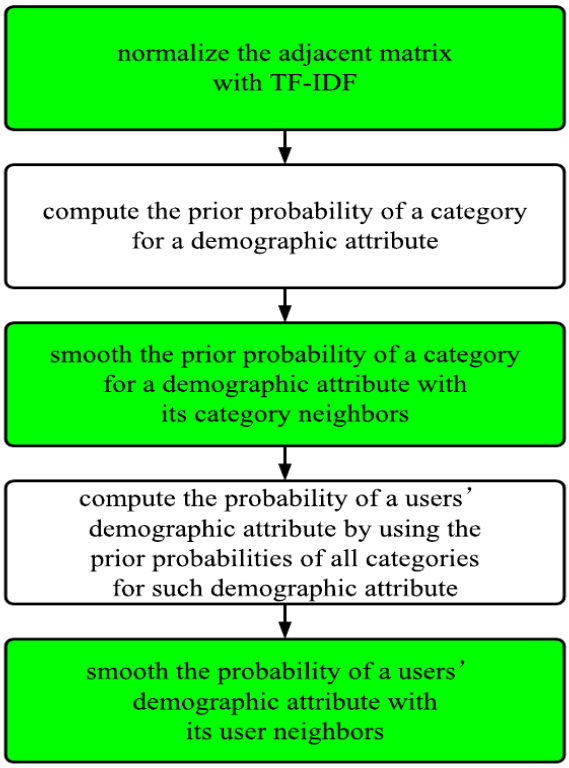


Fig. 5. The Prediction Result Labeled as *Both*

with similar interests. Intuitively, we attempt to optimize the prediction by leveraging Collaborative Filtering (CF) [29] [30], which is sensitive to data sparseness [31]. However, the adjacent matrices R and $R^{'}$ are very sparse, it would import much noise when leveraging the demographic attributes of neighboring users. In many recommendation systems [32] [33], Singular Value Decomposition (SVD) [34] as a matrix factorization algorithm, has been proved to be capable of solving the data sparseness problem. The orthogonal dimensions resulting from SVD are less noisy than the original data. They can capture the latent relationship between users and categories [32]. In this subsection, we use SVD to factorize the adjacent matrix $R^{'}$.

5.4.1 Singular value decomposition

To factorize the adjacent matrix $R^{'}$, we use a new matrix factorization technique called Funk-SVD [35], which factors a $m \times n$ matrix into two matrices as following:

$$R^{'} = R^{'}_{u} * R^{'H}_{c} \tag{10}$$

where, $R^{'}_{u} \in R^{'m \times k}$ and $R^{'}_{c} \in R^{'n \times k}$ are the feedback matrices of users and categories, k is the dimension of SVD. It defines a cost function with the evaluation criterion of Root Mean Squared Error (RMSE) and uses Stochastic Gradient Descent [36] to learn from the training data to get the feedback matrices, which is an iterative procedure. The iterations and dimension k of SVD are parameters in our paper. Theoretical details can be found in [32] and [35]. Based on the users' feedback matrix $R^{'}_{u}$ and categories' feedback matrix $R^{'}_{c}$, we compute the similarity of users and categories to select their neighbors, and then leverage the demographic attributes of their neighbors to smooth the prediction. The similarity of two categories are computed according to Pearson Correlation Coefficient [37]:

$$sim(i,j) = \frac{\sum_{t=1}^{k}(R^{'}_{ik} - \overline{R^{'}_{i}})(R^{'}_{jk} - \overline{R^{'}_{j}})}{\sqrt{\sum_{t=1}^{k}(R^{'}_{ik} - \overline{R^{'}_{i}})^2 \sum_{t=1}^{k}(R^{'}_{jk} - \overline{R^{'}_{j}})^2}} \tag{11}$$

where, $\overline{R^{'}_{i}}$ is the average of the i^{th} row vector $R^{'}_{i}$ in matrix $R^{'}_{c}$, and k is the dimension of $R^{'}_{i}$. The similarity of two users are computed in a similar way.

5.4.2 Smooth the prediction with category neighbors

By computing the similarity between categories, we can identify N neighbors, which are most similar to the i^{th} category based on categories' feedback matrix $R^{'}_{c}$. Then, we denote the top N similar categories of the i^{th} category as a neighbor set $R^{N}_{(c,i)}$. The category i's neighbor prior probability of demographic attribute A is denoted as below:

$$Pr(R^{'N}_{(c,i)}|A) = \frac{1}{N} \sum_{c_j \in R^{N}_{(c,i)}} sim(i,j) Pr(c_j|A) \tag{12}$$

where, $Pr(c_i|A)$ is the original prior probability of the demographic attribute A, $sim(i,j)$ is the similarity value between the i^{th} category and the j^{th} category. Thus, the prior probability of demographic attribute A for category i is smoothed as below:

$$Pr(c_i|A)_{smoooth} = \alpha Pr(c_i|A) + (1-\alpha) Pr(R^{'N}_{(c,i)}|A) \tag{13}$$

where, α is a parameter to control the weight of category i's neighbor prior probability of demographic attribute A.

5.4.3 Smooth the prediction with user neighbors

When we obtain the probability of a user's demographic attribute from (9), we can smooth the prediction with user neighbors. In a similar way as computing category neighbors, we compute the similarity between users and get the top M most similar neighbors of each user u_i based on the users feedback matrix $R^{'}_{u}$. The user i's neighbor probability of demographic attribute A is denoted as below:

$$\Pr(A|R^{'M}_{(u,i)}) = \frac{1}{M} \sum_{u_j \in R^{M}_{(u,i)}} sim(i,j) Pr(A|u_j) \tag{14}$$

where, $R^{'M}_{(u,i)}$ is the set of top M most similar neighbors of user u_i, $sim(i,j)$ is the similarity value between user u_i and user u_j. $Pr(A|u_j)$ is the probability of user u_j's demographic attribute A, where user u_j is a element in set $R^{'M}_{(u,i)}$. Thus, the probability of user u_i's demographic attribute A is smoothed as below:

$$Pr(A|u_i)_{smoooth} = \beta Pr(A|u_i) + (1-\beta) Pr(A|R^{'M}_{(u,i)}) \tag{15}$$

where, β is the parameter to control the weight of user i's neighbor probability of demographic attribute A.

Now, in the proposed approach, we can have four scenarios of prediction results:(i) the prediction result labeled

as *No Smoothing*, which only pre-process the adjacent matrix with TF-IDF; (ii) the prediction result labeled as *Only Category*, which is smoothed by categories neighbors, by importing the result of (13) to (9); (iii) the prediction result labeled as *Only User*, which is smoothed by user neighbors, by importing the result of (9) to (15); and (iv) the prediction result labeled as *Both*, which is the combination of (ii) and (iii) as shown in Fig. 5, smoothed by both categories neighbors and users neighbors.

6 PERFORMANCE EVALUATION

In this section, we evaluate the performance of our approach, including experimental configuration, prediction results and the comparison with baseline prediction.

6.1 Configration

In the experiments, 10-fold cross-validation was deployed for our dataset. The duration of training data lasts 16 weeks. To maximize the prediction results in terms of $F1$ and Acc, we need to estimate five parameters, which are the SVD dimension k when factorizing adjacent matrices R and $R^{'}$, the number of category neighbors N and liner combination parameter α when smoothing the prediction with category neighbors, and the number of user neighbors M and liner combination parameter β when smoothing the prediction with user neighbors.

6.1.1 SVD dimension k

When we estimate the SVD dimension k, we first set other parameters as fixed values (*e.g.*, $\alpha = 0.8$, $\beta = 0.8$, $N = 10$ *and* $M = 10$) [38]. Then, we train the model with different SVD dimension k ranges from 5 to 100 and different SVD iterations ranges from 0 to 100. Fig. 6 shows the prediction performance related with SVD dimension k. The prediction performance is improved with the increase of SVD dimension k. When k reaches 71, a stable prediction performance can be obtained. The values of $F1$ and Acc achieve the best performance of 75.1% and 75.9%, while k is 71 and SVD iteration is 62. Thus, we set k and the value of iteration as 71 and 62 in the following experiments, respectively.

6.1.2 Smooth the prediction with category neighbors

We have deployed experiments to study the influence of categories neighbors on prediction performance. We first set β and M at fixed values (*e.g.*, $\beta = 0.8$, $M = 10$), then attempt to find the best value for the number of neighbors N and parameter α. In the experiments, N ranges from 1 to 120, while α ranges from 0.05 to 1. As shown in Fig. 7 and Fig. 8, the prediction performance is related with the number of category neighbors N and parameter α. We found that both the $F1$ and Acc are improved when N increased. The prediction achieves the best results ($F1$=79.68% and Acc=80.3%) when $N = 86$ and $\alpha = 0.84$.

6.1.3 Smooth the prediction with user neighbors

In a similar way, we tested different values for user neighbors M and parameter β to optimize the prediction. Base on above experimental results, we set $N = 86$ and $\alpha = 0.84$, then tested M from 1 to 150 and β from 0.05 to 1. As shown

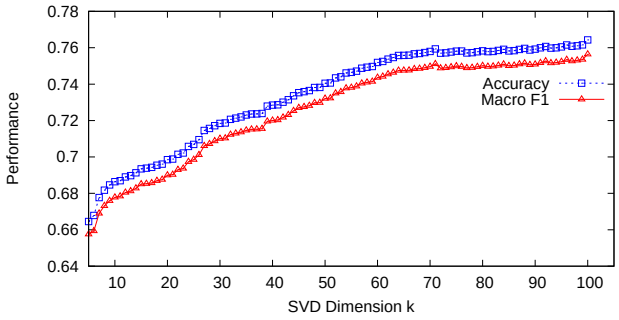


Fig. 6. Performance with SVD dimensions

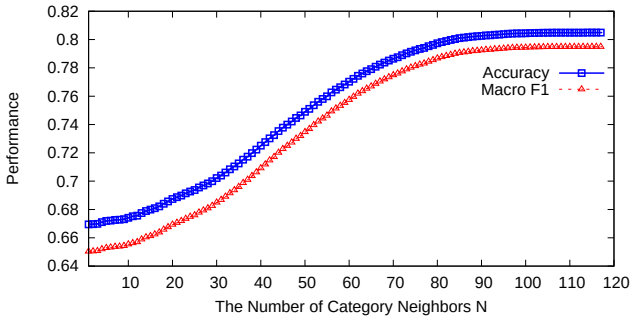


Fig. 7. Performance with different categories neighbors

in Fig. 9 and Fig. 10, the prediction performance is related with the M and β. We found that both the $F1$ and Acc are improved when M increased. The prediction achieves the best results ($F1$=79.4% and Acc=80.2%) when $M = 35$ and $\beta = 0.82$.

6.2 Prediction Results

In this subsection, for the four scenarios of proposed approach, we provide the prediction results in gender, phone level and age group.

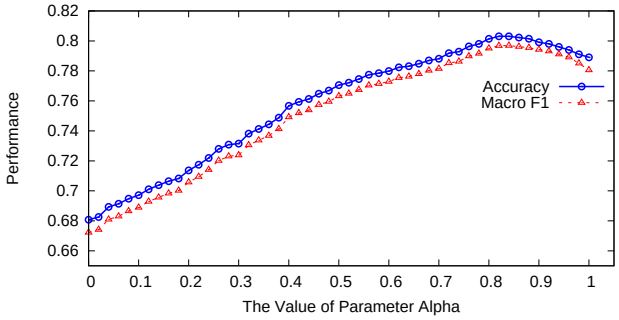


Fig. 8. Performance with different values of α

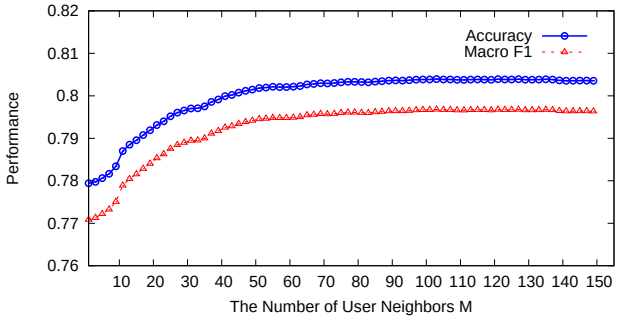


Fig. 9. Performance with different users neighbors

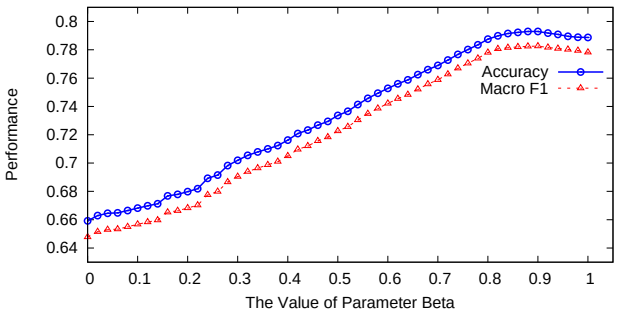


Fig. 10. Performance with different values of β

6.2.1 Prediction results for gender

Table 3 shows the results of gender prediction. In particular, without any smoothing (labeled as *No Smoothing*), our approach for gender prediction can achieve 68.34% and 69.50% in terms of $F1$ and Acc. As shown in Table 3, our smoothing methods are capable of improving the prediction results. If we smooth the prediction results with either category neighbors or user neighbors (labeled as *Only Category* and *Only User*), the value of $F1$ and Acc can be improved to around 75%. Moreover, if both category neighbors and user neighbors as used to smooth the prediction results (labeled as *Both*), we can receive the best results at 80.11% and 81.21% in terms of $F1$ and Acc.

TABLE 3
Smooth the Prediction Results for Gender

		Rec	Prec	Micro F1	Macro F1	Acc
No Smoothing	Male	68.82%	80.95%	74.4.%	68.34%	69.50%
	Female	70.73%	55.65%	62.29%		
Only Category	Male	75.72%	85.68%	80.39%	75.21%	76.29%
	Female	77.32%	63.98%	70.02%		
Only User	Male	73.32%	85.24%	78.83%	73.80%	74.76%
	Female	77.33%	61.90%	68.76%		
Both	Male	81.75%	88.08%	84.80%	80.11%	81.21%
	Female	80.25%	71.13%	75.42%		

TABLE 4
Smooth the Prediction Results for Phone Level

		Rec	Prec	Micro F1	Macro F1	Acc
No Smoothing	Level1	70.28%	54.23%	61.22%	52.55%	55.51%
	Level2	58.49%	57%	57.74%		
	Level3	29.51%	56%	38.71%		
Only Category	Level1	61.86%	75.79%	68.15%	64.28%	65.32%
	Level2	69.05%	59.33%	63.82%		
	Level3	67.86%	55.20%	60.88%		
Only User	Level1	55.51%	73.75%	63.34%	57.01%	59%
	Level2	60.45%	58.42%	59.42%		
	Level3	67.86%	37.46%	48.27%		
Both	Level1	84.43%	72.67%	78.11%	71.82%	73.85%
	Level2	69.64%	75.07%	72.25%		
	Level3	57.61%	74.84%	65.11%		

6.2.2 Prediction results for phone level

Table 4 shows the results of Phone Level prediction, and proves that our smoothing methods effectively improved the prediction results. In particular, without any smoothing, our approach for Phone Level prediction can achieve 52.55% and 55.51% in terms of $F1$ and Acc. If we smooth the prediction results with user neighbors, the value of $F1$ and Acc can be improved to around 60%. If we smooth the prediction results with category neighbors, the value of $F1$ and Acc can be improved to around 65%. Moreover, if both category neighbors and user neighbors are used to smooth the prediction results, we can get the best results at 71.82% and 73.85% in terms of $F1$ and Acc.

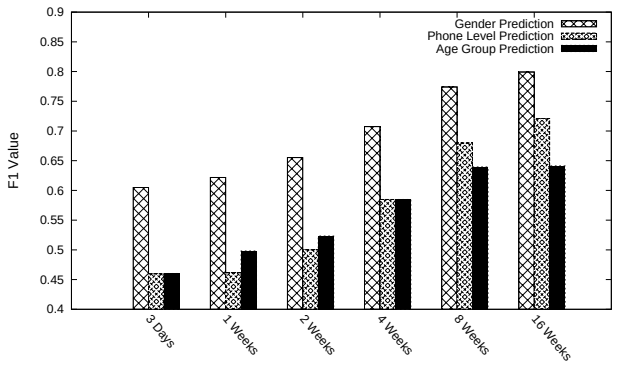
6.2.3 Prediction results for age group

TABLE 5
Smooth the Prediction Results for Age Group

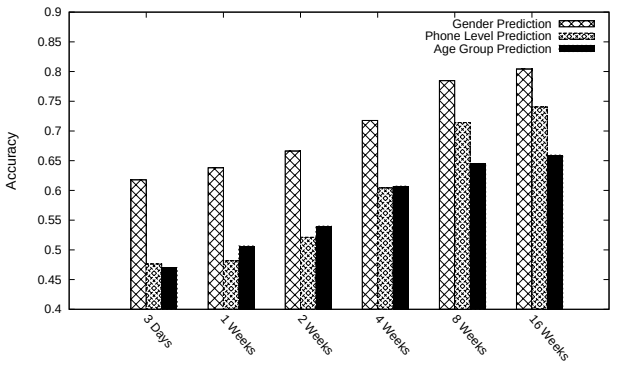
		Rec	Prec	Micro F1	Macro F1	Acc
No Smoothing	Teenage	56.80%	37.75%	45.35%	50.95%	54.89%
	Youngster	53.63%	47.76%	50.52%		
	Young	48.49%	41.62%	44.8%		
	Mid-age	67.06%	61.74%	64.29%		
	Elder	40.54%	64.45%	49.77%		
Only Category	Teenage	63.92%	50.85%	56.64%	59.98%	63.89%
	Youngster	61.24%	56.5%	58.78%		
	Young	56.36%	62.49%	59.26%		
	Mid-age	77.43%	69.40%	73.19%		
	Elder	47.56%	57.39%	52.02%		
Only User	Teenage	53.11%	32.50%	40.33%	56.07%	61.87%
	Youngster	61%	62.33%	61.66%		
	Young	62.15%	51.44%	56.29%		
	Mid-age	69.63%	69.40%	69.52%		
	Elder	45.91%	61.42%	52.55%		
Both	Teenage	69.02%	63.96%	66.4%	64.39%	66.42%
	Youngster	64.50%	62.32%	63.36%		
	Young	69.63%	61.26	65.03%		
	Mid-age	72.74%	72.19%	72.46%		
	Elder	49.26%	61.42%	54.67%		

Table 5 shows the results of Age Group prediction, and also proves that our smoothing methods effectively improved the prediction results. Similarly, without any smoothing, our approach for Age Group prediction can achieve 50.90% and 54.89% in terms of $F1$ and Acc. If we smooth the prediction results with category neighbors, the $F1$ and Acc can be improved to 59.98% and 63.89%. If we smooth the prediction results with user neighbors, the $F1$ and Acc can be improved to around 60%. If both category neighbors and user neighbors as used to smooth the prediction results, we can receive the best results at 64.39% and 66.42% in terms of $F1$ and Acc.

All the above experimental results show that our prediction approach can achieve good performance and the proposed smoothing methods effectively improve the prediction results. It can be found that only category smoothing outperform only user smoothing. Moreover, the prediction approach achieved the best performance by smoothing results with both category neighbors and user neighbors. In this way, for gender prediction, we get the best results at 80.11% and 81.21% in terms of $F1$ and Acc. Similarly, the best values of $F1$ and Acc for predicting users' phone level are 71.82% and 73.85%. For the more challenging multi-class problem, the prediction of users' age group is to classify users into one of five candidate sets, where the prediction achieves 64.39% and 66.42% in terms of $F1$ and Acc. Compared with the results of gender prediction and phone level prediction, there is a little backslide in the prediction of age group, especially in the prediction of elder people. The

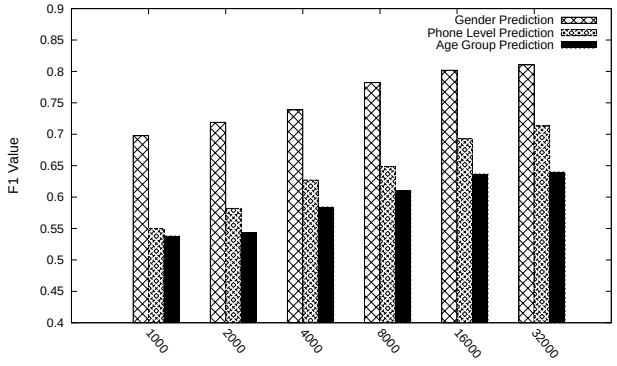


(a) F1 Value in Different Data Duration Prediction

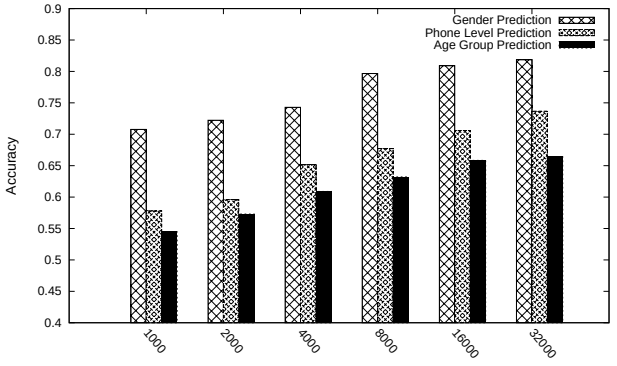


(b) Accuracy in Different Data Duration

Fig. 11. Prediction Performance in Different Data Duration



(a) F1 Value in Different User Amount



(b) Accuracy in Different User Amount

Fig. 12. Prediction Performance in Different User Amount

TABLE 6
Comparison in Prediction with Different Approaches

		Logistic	NB	DT	SVM	Our
Gender	Acc	62.29%	64.85%	65.23%	72.31%	81.21%
	Macro F1	62.13%	63.41%	64.51%	71.05%	80.11%
Phone Level	Acc	59.81%	45.34%	59.25%	64.03%	73.84%
	Macro F1	57.16%	44.75%	58.12%	62.85%	71.82%
Age Group	Acc	53.14%	45.47%	52.43%	59.14%	66.42%
	Macro F1	50.15%	44.15%	50.05%	56.93%	64.39%

reason is two-fold. On one hand, the prediction of age group is a more challenging multi-class problem. On the other hand, the correlation between age group and categories could be weaker than the correlation between gender (or phone level) and categories.

6.3 Comparison With Baselines

Here, by using the same dataset and the same duration of training data, we compare the performance of proposed approach with the baselines stated in Section 4 (*e.g.*, Logistic, NB, DT, and SVM). As shown in the Table 6, our approach outperforms the four classifiers. Relatively, in term of $F1$ and Acc, the improvements on gender prediction are 19.08% and 17.82% over Logistic, 17.8% and 15.26% over NB, 16.70% and 14.88% over DT,10.16% and 7.81% over SVM, respectively. For the prediction of users' phone level, the improvements are 16.69% and 12.02% over Logistic, 29.1% and 26.48% over NB, 15.72% and 12.57% over DT, 10.99% and 7.79% over SVM in term of $F1$ and Acc, respectively. For predicting users' age group, in term of $F1$ and Acc, the improvements are is 16.27% and 11.24% over Logistic, 22.3% and 18.92% over NB, 16.37% and 11.96% over DT, 9.5% and 5.25% over SVM, respectively.

It should be noted that our approach is based on NB which has the worst performance that 63.41% in $F1$ and 64.85% in Acc for gender prediction, 44.75% in $F1$ and 45.34% in Acc for phone level prediction, 44.15% in $F1$ and 45.47% in Acc for age group prediction, respectively. Our approach improves the performance by dealing with category-unbalance problem and smoothing results with user/category neighbors. The *No Smoothing* cases in Table 3, Table 4 and Table 5 have shown the effect of dealing with category-unbalance problem from which the improvements are 4.93% in $F1$ and 4.65% in Acc for gender prediction, 7.8% in $F1$ and 10.17% in Acc for phone level prediction, 6.8% in $F1$ and 9.42% in Acc for age group prediction, respectively. The *Both* cases in Table 3, Table 4 and Table 5 have shown the effect of smoothing results with user/category neighbors from which the further improvements are 11.77% in $F1$ and 11.71% in Acc for gender prediction, 19.27% in $F1$ and 18.34% in Acc for phone level prediction, 13.44% in $F1$ and 11.53% in Acc for age group prediction, respectively. The results verify that both the way of dealing with category-unbalance problem and the optimizing method of smoothing results with user/category neighbors are effective. By the way of contrast, the optimizing method

of smoothing results with user/category neighbors is more important.

Next, we compare the *convergent property* of the proposed approach with that of the baselines. In particular, by varying the data duration and user amount, we evaluate the performance of proposed approach, and analyzed the required data duration and user amount for the performance to achieve a stable status.

By using the fixed users amount of 32,000 users (the same as the experiments state in Section 4), Fig. 11 shows the prediction performance of the proposed approach in different data duration. It can be noticed that, with the increase of data duration, both $F1$ and Acc value increase dramatically while the data duration varies from 3 days to 8 weeks. When the data duration lasts more than 8 weeks, the prediction performance achieves a stable status with subtle improvement. For the baselines, as shown in Fig. 1, their prediction results in $F1$ and Acc, achieve stable status with the duration of more than 8 weeks. It means that, if we just consider the required data duration, the *convergent property* of the proposed approach is the same as that of the baselines.

Similarly, by setting the data duration as 16 weeks, Fig. 12 shows the prediction performance of the proposed approach in different user amount. It can also be noticed that, with the increase of user amount, both $F1$ and Acc value increase dramatically while the user amount varies from 1,000 to 8,000. When the user amount is more than 8,000, the prediction performance achieves a stable status with subtle improvement. For the baselines, as shown in Fig. 2, their prediction results in $F1$ and Acc, achieve stable status with the user amount of more 4,000. It means that, when we consider the required user amount, the *convergent property* of the baselines is better than that of the proposed approach. This is because the proposed approach need to compute the information of user neighbors, and further leverage it to smooth the prediction results.

Although the proposed approach required more user amount to enable its prediction results to achieve stable status, it should be noted that its performance still outperforms the baselines with the same user amount before it reaches its convergent status (*e.g.*, at the user amount of 4,000, the baselines achieve stable status, while the proposed approach does not). As shown in Fig. 12, at the user amount of 4,000, the performance of proposed approach is 73.89% and 74.28% in terms of $F1$ and Acc for gender prediction, 62.71% and 65.15% in terms of $F1$ and Acc for phone level prediction, and 58.42% and 60.9% in terms of $F1$ and Acc for age group prediction. While as shown in Fig. 2, at the user amount of 4,000, the performance of the best case (*e.g.*, SVM) in baselines is 69.87% and 71.38% in terms of $F1$ and Acc for gender prediction, 61.98% and 63.89% in terms of $F1$ and Acc for phone level prediction, and 55.92% and 57.14% in terms of $F1$ and Acc for age group prediction. Table 7 shows the detailed comparison at user amount of 4,000.

Moreover, we have conducted some other prediction of education attainment and personal income. Education attainment prediction, whose performance is more than 83% in both $F1$ and Acc, is a binary classification problem to predict if a user is a college graduate or not. Personal income prediction, whose performance is more than 70% in both $F1$ and Acc, is a four classification problem to classify a user into one of four candidate group.

TABLE 7
Performance Comparison at the User Amount of 4,000

		Logistic	NB	DT	SVM	Our
Gender	Acc	61.29%	62.14%	63.15%	71.38%	74.28%
	Macro F1	60.23%	61.87%	63.71%	69.87%	73.89%
Phone Level	Acc	57.61%	45.04%	58.75%	63.89%	65.15%
	Macro F1	56.06%	44.57%	57.72%	61.98%	62.71%
Age Group	Acc	52.64%	45.23%	52.03%	57.14%	60.9%
	Macro F1	49.14%	43.11%	49.65%	55.92%	58.42%

7 Discussion

We firstly want to discuss the computation efficiency of proposed approach. As it adds the processes of dealing with category-unbalance problem and smoothing results with user/category neighbors, it consumes more computation time that indicates a trade off between computation efficiency and accuracy. Fortunately, as it only leverages the top M user neighbors and top N category neighbors for prediction results optimization, it can cope well with the increased dataset size. Next, we want to state that there is a possibility for further improving prediction accuracy. The experimental parameters configuration is a five-variable optimization problem. In our configuration, it can receive good results but might not be the best, mathematically speaking.

8 Conclusions

This paper proposed a novel approach to predict demographic information by leveraging the perspectives of smartphone application usage. Our contributions are as follows. Firstly, we proposed to match the entries in smartphone applications log with types of categories according to their topics, and build a matrix to correlate users' demographic attributes with the categories. Secondly, in order to avoid over weighted categories biasing prediction results, we proposed to process the matrix and deal with the category-unbalance problem before importing it to Bayes-based classifier. Thirdly, we provided a method to optimize the prediction results by smoothing the prediction results with category neighbors and user neighbors. The experimental results show that, our approach outperforms the baselines, and achieves 80.11% and 81.21% of $F1$ and Acc in gender prediction, 71.82% and 73.84% of $F1$ and Acc in phone level prediction, and 64.39% and 66.42% of $F1$ and Acc in age group prediction, respectively.

9 Acknowledgments

This work was supported in part by the National Science Foundation of China (No. 61133016, No. 61300191, No. 61202445 and No. 61370026), the National High Technology Research and Development Program of China (No. 2015AA016007), the Sichuan Science-Technology Support Plan Program (No. 2014GZ0106 and No. 2016JZ0020), the Fundamental Research Funds for the Central Universities (No. 2014SCU11013), the Canadian Natural Sciences and Engineering Research (NSERC), the NSERC DIVA Strategic Research Network, and various industry partners.

REFERENCES

[1] A. Hannak, P. Sapiezynski, A. M. Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson, "Measuring personalization of web search," in *WWW '13*.

[2] E. G. Smit, G. V. Noort, and H. A. Voorveld, "Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in europe," *Computers in Human Behavior*, 2014.

[3] B. J. Jansen, K. Moore, and S. Carman, "Evaluating the performance of demographic targeting using gender in sponsored search," *Information Processing and Management*, 2013.

[4] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing User Behavior in Mobile Internet," *IEEE Transactions on Emerging Topics on Computing*, vol. 3, no. 1, pp. 95–106, 2015.

[5] G. Nikesh and Y. Yarowsky, "Modeling latent biographic attributes in conversational genres," in *ACL '09*.

[6] X. Yan and L. Yan, "Gender classificaiton of weblog authors," in *AAAI '06*.

[7] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in *EMNLP '11*.

[8] D. Kelly, S. Brendan, and C. Brian, "Uncovering Measurements of Social and Demographic Behavior From Smartphone Location Data," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 2, pp. 188–198, 2013.

[9] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classificaiton," in *AAAI '11*.

[10] O. Jahna, "Inferring gender ofmovie reviewers: Exploitingwriting style, content and metadata," in *CIKM '10*.

[11] I. Weber and C. Castillo, "The demographics of web search," in *SIGIR '10*.

[12] I. Weber and A. Jaimes, "Demographic information flows," in *CIKM '10*.

[13] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on users browsing behavior," in *WWW '07*.

[14] K. Santosh, E.-H. Han, and G. Karypis, "Content-based methods for predicting web-site demographic attributes," in *ICDM '10*.

[15] S. Goel, M. H. Jake, and M. I. Sirer, "Who does what on the web: A large-scale study of browsing behavior," in *ICWSM '12*.

[16] S. Seneviratne, A. Seneviratne, P. Mohapatra, and A. Mahanti, "Predicting user traits from a snapshot of apps installed on a smartphone," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 18, no. 2, pp. 1–8, 2014.

[17] Z. Qin, Y. Wang, Y. Xia, H. Cheng, Y. Zhou, Z. Sheng, and V. Leung, "Demographic information prediction based on smartphone application usage," in *SMARTCOMP '14*.

[18] "PhonePrice Range," http://detail.zol.com.cn/cell_phone_index/subcate57_list_1.html.

[19] "2010 sixth national census data bulletin of China," http://www.gov.cn/test/2012-04/20/content_2118413.htm.

[20] N. Deng, Y. Tian, and C. Zhang, *Support vector machines: optimization based theory, algorithms, and extensions*, 2012.

[21] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, "Applied logistic regression," *John Wiley & Sons*, 2013.

[22] C. Bielza and P. Larraaga, "Discrete bayesian network classifiers: A survey," *ACM Computing Surveys (CSUR)*, 2014.

[23] R. C. Barros, M. P. Basgalupp, A. De Carvalho, and A. A. Freitas, "A survey of evolutionary algorithms for decision-tree induction," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 2012.

[24] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *PNAS 2013*.

[25] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML ' 1997*.

[26] P. Soucy and G. W. Mineau, "Beyond tfidf weighting for text categorization in the vector space model," in *IJCAI '2005*.

[27] J. Leskovec, R. Anand, and D. U. Jeffrey, "Mining of massive datasets," *Cambridge University Press*, 2014.

[28] T. D. Nielsen and F. V. JENSEN, "Bayesian networks and decision graphs," *Springer Science & Business Media*, 2009.

[29] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *UAI 1998'*.

[30] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, 2009.

[31] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of dimensionality reduction in recommender system-a case study," DTIC Document, Tech. Rep., 2000.

[32] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *SIGKDD '2008*.

[33] M. H. Pryor, "The effects of singular value decomposition on collaborative filtering," 1998.

[34] G. H. Golub and C. F. Van Loan, *Matrix computations*, 2012.

[35] "FunkSVD," http://sifter.org/~simon/journal/20061211.html.

[36] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *COMPSTAT '2010*.

[37] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *WWW '2001*.

[38] P. Hall, B. U. Park, and R. J. Samworth, "Choice of neighbor order in nearest-neighbor classification," *The Annals of Statistics*, 2008.

Zhen Qin is currently an associate professor in the School of Information and Software Engineering, University of Electronic Science and Technology of China (UESTC). He received his Ph.D. degree from UESTC in 2012. He was a visiting scholar in the Department of Electrical Engineering and Computer Science at Northwestern University. His research interests include network measurement, mobile social networks and wireless sensor networks.

Yilei Wang is currently a Ph.D. student in the School of Information and Software Engineering, University of Electronic Science and Technology of China (UESTC). His research interests include mobile Internet and data mining.

Hongrong Cheng is currently an associate professor in the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC). She received her Ph.D. degree from UESTC in 2009. Her research interests include mobile social networks, machine learning and data mining.

Yingjie Zhou is currently an assistant professor in the College of Computer Science at Sichuan University (SCU), China. He received his Ph.D. degree from University of Electronic Science and Technology of China (UESTC) in 2013. He was a visiting scholar in the Department of Electrical Engineering at Columbia University. His current research interests include network management, resource allocation and behavioral data analysis in computer networks, smart grids.

Zhengguo Sheng is currently a Lecturer with Department of Engineering and Design, University of Sussex. He received his Ph.D. degree from Imperial College London in 2011. His current research interests include IoT/M2M, cloud/edge computing, vehicular communications, and power line communication. Dr. Sheng received IEEE Outstanding Service Award 2015, the Auto21 TestDRIVE Competition Award in 2014 and the Orange Outstanding Researcher Award in 2012.

Victor C. M. Leung is currently a Professor of electrical and computer engineering and the holder of the TELUS Mobility Research Chair with the University of British Columbia (UBC). He received his Ph.D. degree in electrical engineering from UBC in 1982. His research interests include wireless networks and mobile systems, where he has coauthored more than 700 technical papers in archival journals and refereed conference proceedings, several of which having won best paper awards. Dr. Leung is a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada.